From Big Data Analytics
To Smart Data Analytics
With Parallelization Techniques

Dr. – Ing. Morris Riedel et al.

Adjunct Associated Professor, University of Iceland

Jülich Supercomputing Centre, Germany

Head of Research Group High Productivity Data Processing

IEEE Talk, University of Iceland, 16th September 2014

HELMHOLTZ ASSOCIATION

Research Field Key Technologies

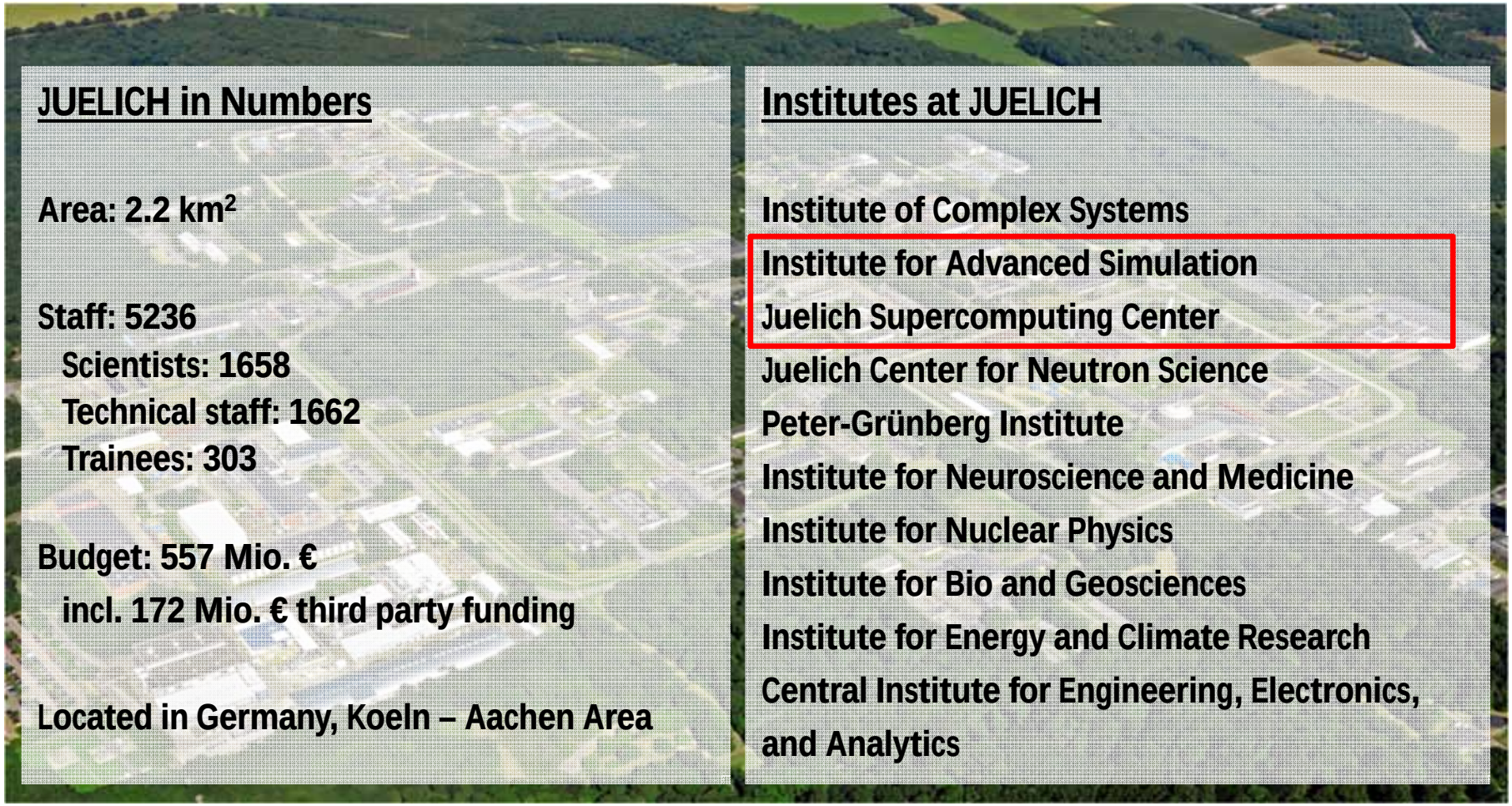
Jülich Supercomputing Centre

Supercomputing & Big Data

JÜLICH FORSCHUNGSZENTRUM

UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

Research Centre Juelich – Forschungszentrum Juelich



JUELICH in Numbers

Area: 2.2 km²

Staff: 5236
- Scientists: 1658
- Technical staff: 1662
- Trainees: 303

Budget: 557 Mio. €
- incl. 172 Mio. € third party funding

Located in Germany, Koeln – Aachen Area

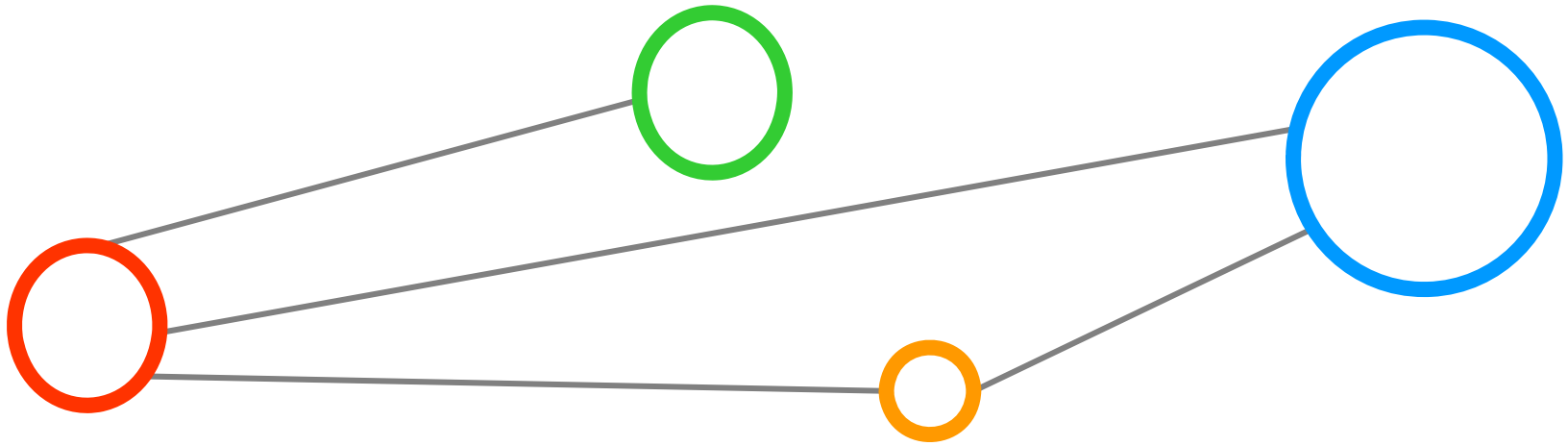
Institutes at JUELICH

Institute of Complex Systems
Institute for Advanced Simulation
Juelich Supercomputing Center
Juelich Center for Neutron Science
Peter-Grünberg Institute
Institute for Neuroscience and Medicine
Institute for Nuclear Physics
Institute for Bio and Geosciences
Institute for Energy and Climate Research
Central Institute for Engineering, Electronics, and Analytics

Research for generic key technologies of the next generation → towards a 2020 Vision

Scientific & Engineering Application-driven Problem Solving

Outline



Outline

‘Big Data‘ Challenges

- Demand for Smart Data Analytics
- Understanding Scientific Big Data Analytics
- Recent Technology Advances & Research

Parallel and Scalable Methods

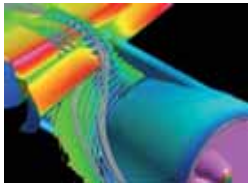
- Parallelization Fundamentals & Paradigms
- Reliable Computing & Data Infrastructures
- Parallel and Scalable Approaches & Tools

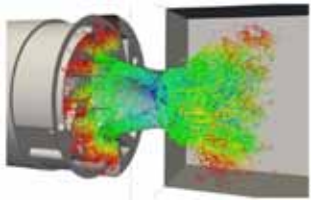
Selected ‘Smart Data Analytics‘ Applications

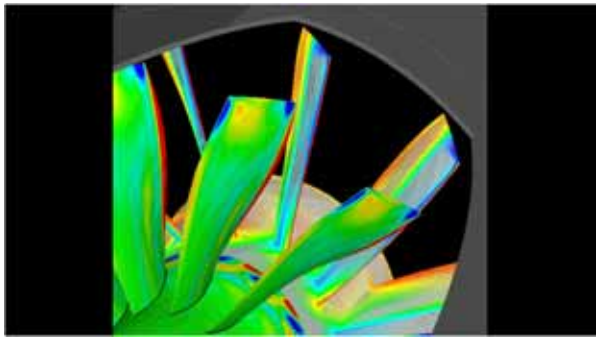
- Ultrascan Bio-Chemical Methods
- Location-based Social Network-based Health Analytics
- Classification methods in Remote Sensing
- Techniques to understand the Human Brain

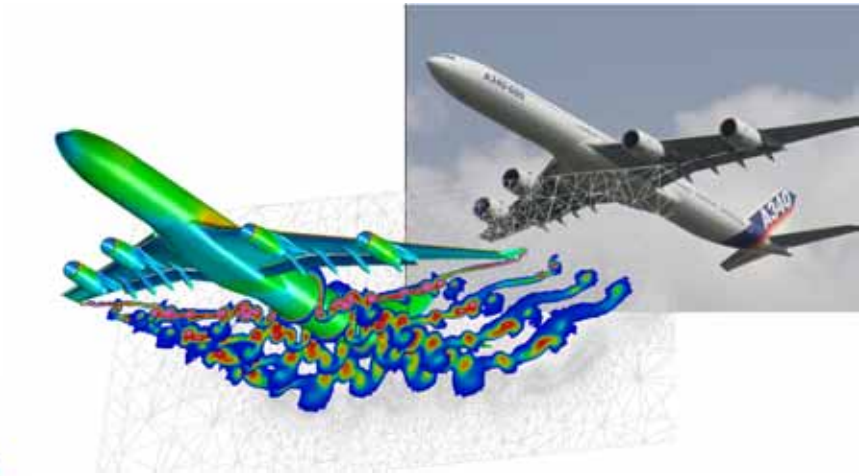
Summary & References

Computational Science & Engineering – increasing ‘Big Data‘



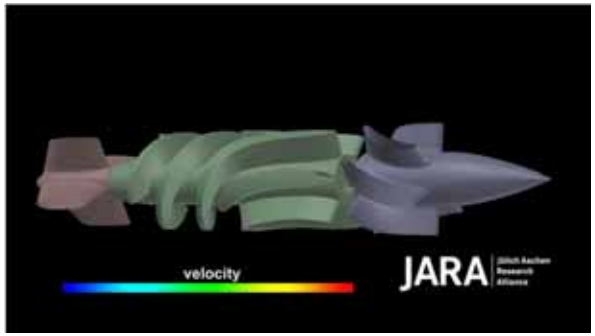


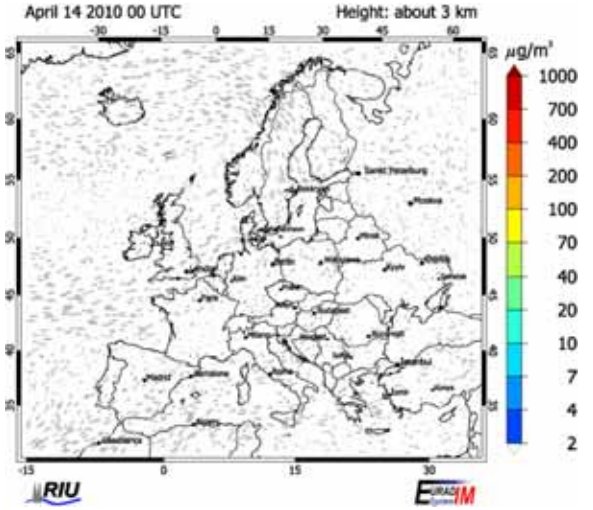




Floating Point Operations per Second (FLOPS)

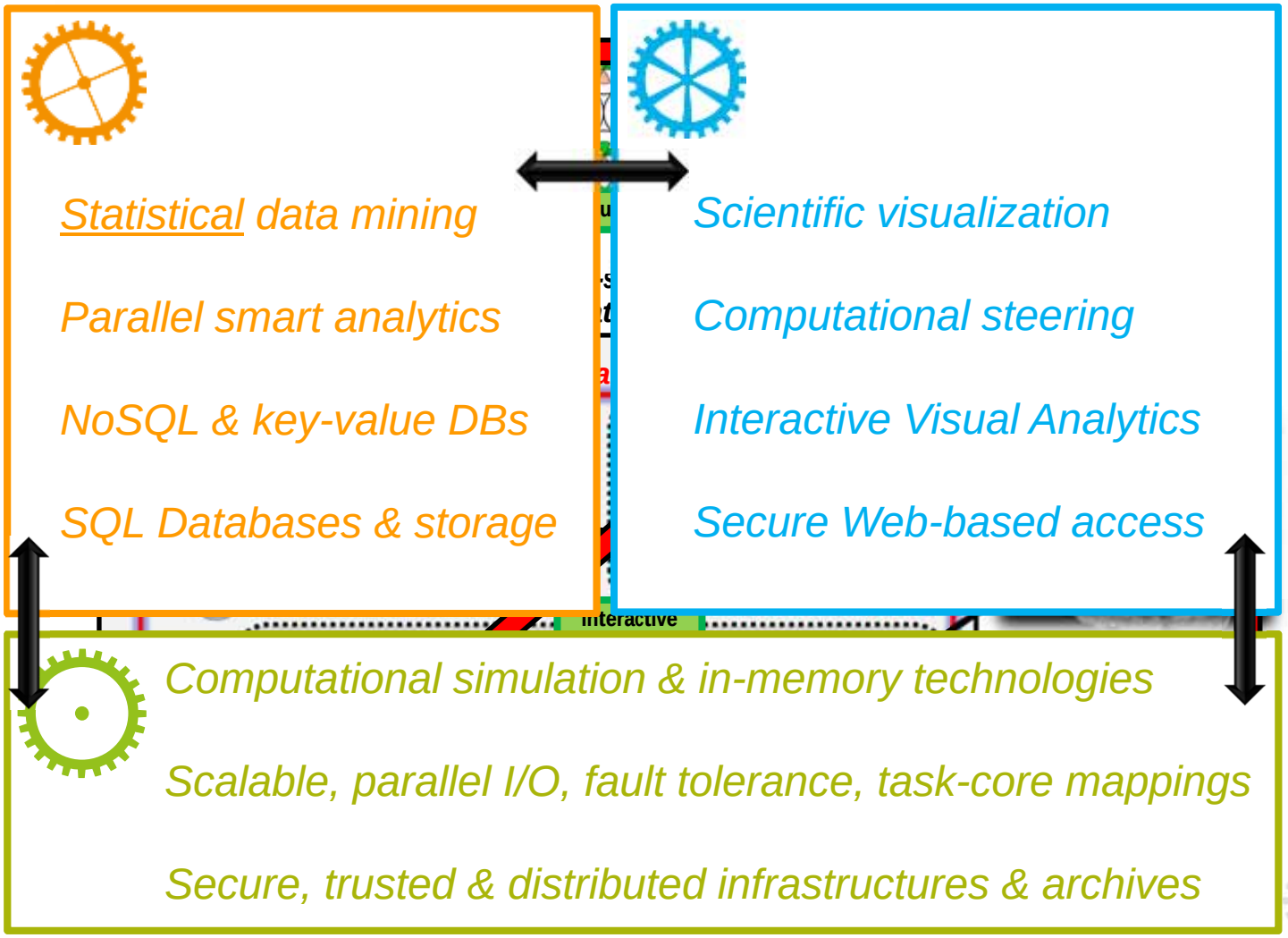
Flops = 10^0
MFlops = 10^6
GFlops = 10^9
TFlops = 10^{12}
PFlops = 10^{15}
EFlops = 10^{18}





- **Use parallelization for computational simulations is mainstream, but growing demands**

Exascale Lighthouse: Strategic Research towards a 2020 Vision

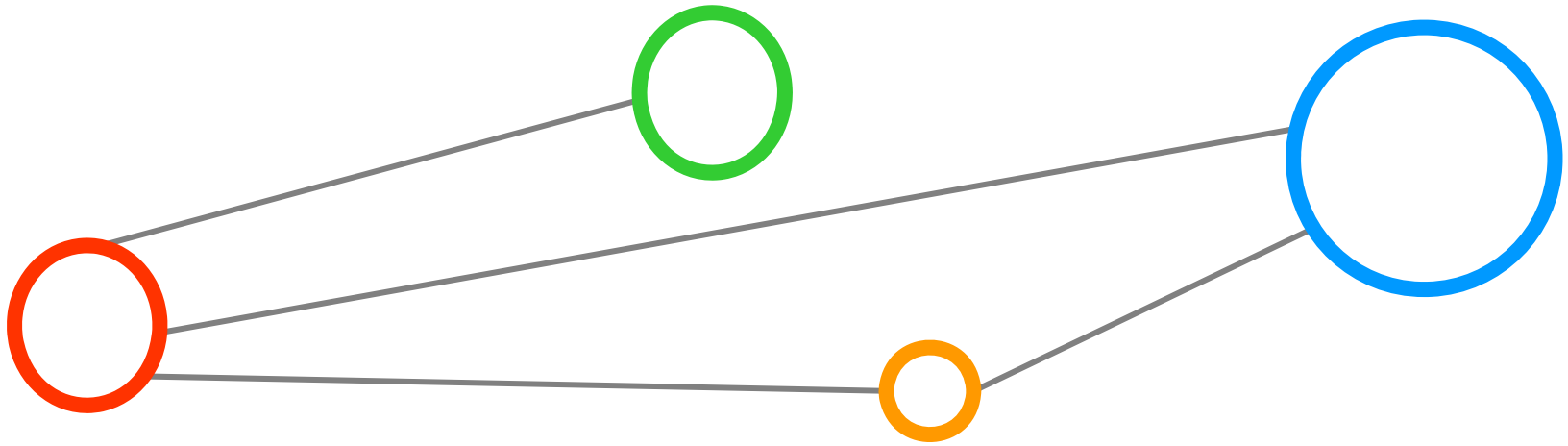


Smart Data Analytics:

Combination of these (often disjunct) Research Areas

Strategic Research towards a 2020 Vision

‘Big Data‘ Challenges



Many Strategic Reports that describe potential of 'Big Data'

'Understanding climate change, finding alternative energy sources, and preserving the health of an ageing population are all cross-disciplinary problems that require high-performance data storage, ***smart analytics****, transmission and mining to solve.'*

[1] Riding the Wave, EC Report, 2010

'In the data-intensive scientific world, ***new skills are needed for*** *creating, handling,* ***manipulating, analysing,*** *and making available large amounts of data for re-use by others.'*

[2] A Surfboard for riding the wave, Report, 2012

[3] DoE ASCAC Report, 2013

'Integration of data analytics *with exascale simulations represents a new kind of workflow that will impact both* ***data-intensive science and exascale computing.'***

Chances and Pitfalls for ‘Scientific Big Data Analytics‘

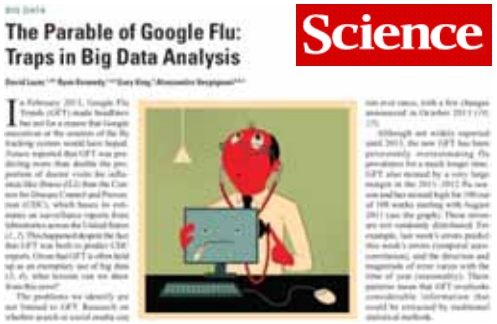
~2009 – H1N1 Virus Made Headlines

- Nature paper from Google employees
- Explains how Google is able to predict fast winter flus
- Not only on national scale, but down to regions
- Possible via logged big data – ‘search queries‘

[4] Jeremy Ginsburg et al., ‘Detecting influenza epidemics using search engine query data’, Nature 457, 2009

~2014 – The Parable of Google Flu

- Large errors in flu prediction & lessons learned
- (1) Dataset: Transparency & replicability impossible
- (2) Study the algorithm since they keep changing
- (3) It‘s not just about size of the data



[5] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, ‘The Parable of Google Flu: Traps in Big Data Analysis’, Science Vol (343), 2014

- **Big data is not always better data – Think about difference of causality vs. correlation**

Academia and Industry share ‘Big Data Challenges‘

Smart Data Innovation Lab

- **Take advantage of cloud technologies & distributed computing/storage**
- **Focus in key areas of science & engineering**

Smart Data Innovation Lab: Human Eye Analytics Example

Computational simulation & in-memory technologies

Smart Data Innovation Lab

[6] SDIL Online, 2014

JÜLICH FORSCHUNGSZENTRUM
Forschungszentrum Jülich GmbH

BAYER
Bayer Technology Services GmbH

Selected Scientific Application Case

- Understand prescriptions, diagnosis and results as longitudinal study
- Take advantage of historic datasets with focus on the human eye treatments
- Data provided by the Munich University Hospital – Human Eyeclinic

Investigation in technologies

- SAP Hana-DB – in-memory database with a focus on analytics
- Software AG Terracotta – in-memory technology with many capabilities

➢ **Research activities with PhD Candidate Christian Bodenstein – Juelich Supercomputing Centre**

Process Data Analytics – ‘Big Data‘ Impact in Engineering

Engineering systems

- Processes, units, and equipment are designed with clear objectives
- Operated under well-controlled circumstances as designed
- Created often from engineers following long traditions (e.g. gas turbine)
- Industry 4.0: Can a factory re-create itself, but better – more efficient?

Potential ‘Big Data‘ Chances

- Take advantage of ‘inferential sensors‘ (field of process systems engineering)
- Historical and real-time data are valuable for safe and efficient operations
- Understanding abnormal process behaviour
- Useful for circumstances that are not considered in the design phase

[7] S. Joe Qin, ‘Process Data Analytics in the Era of Big Data’, AICHE Journal, 2014

- **‘Big Data‘ is not the answer to everything, but useful considering ‘out of design phases‘**

Generic Data Methods Demand in Science & Engineering

Scientific Big Data Analytics

- Advancing user-centered data mining methods & tools (e.g. outlier detection, support vector machines, etc.)

Algorithms and Data Structures

- Enhancing/parallizing generic techniques and algorithmic cores (e.g. indexing, sorting, new DBs)

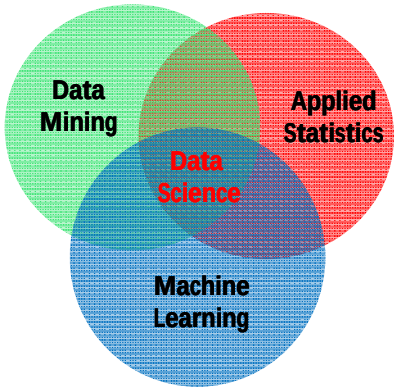
Data Visualization

- Scaling visualization techniques for in-situ high-performance graphics and data visualization

Data Security in Distributed Systems

- Efficient security and privacy solutions applied to real-world big data use-cases, multi-user systems

Research Group High Productivity Data Processing Focus



Scientific Computing

- Scientific Applications using 'Big Data'
- Traditional Scientific Computing Methods
- Different parallelization paradigms
- Smart Data Analytics Approaches & Techniques
- Optimized Data Access & Management
- Statistical Data Mining & Machine Learning

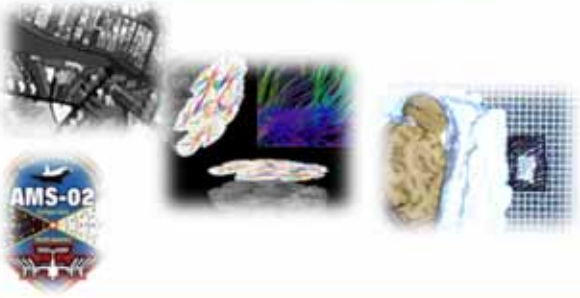
Different Approaches

'Big Data' Methods

Scientific Applications

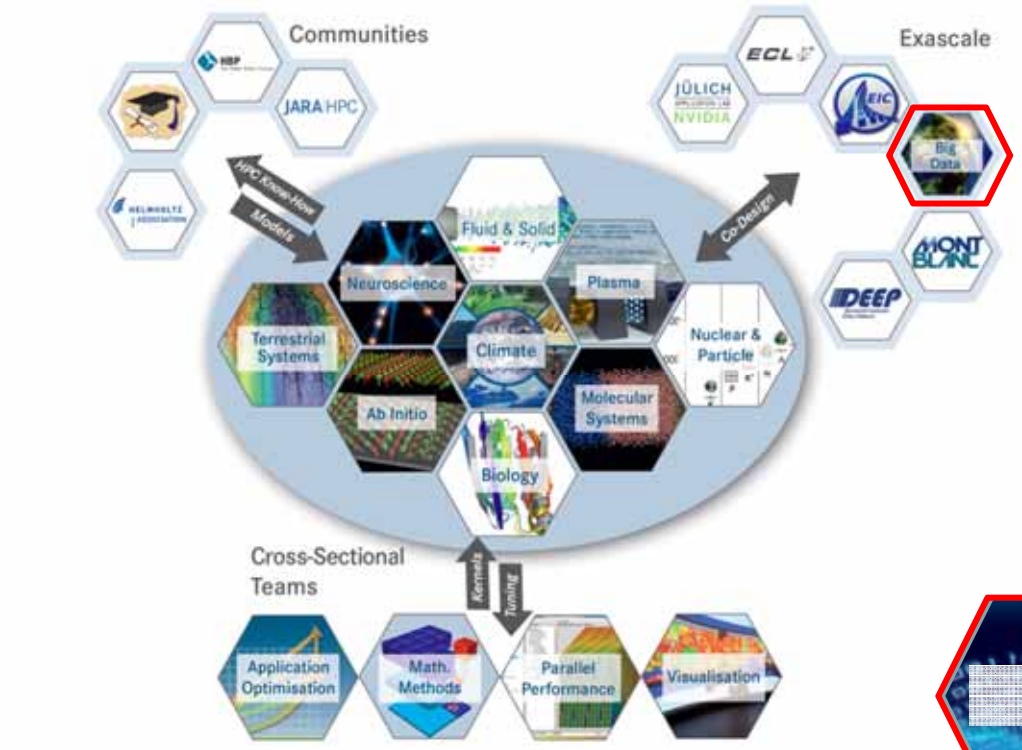
(Parallel) Technologies

Research Data

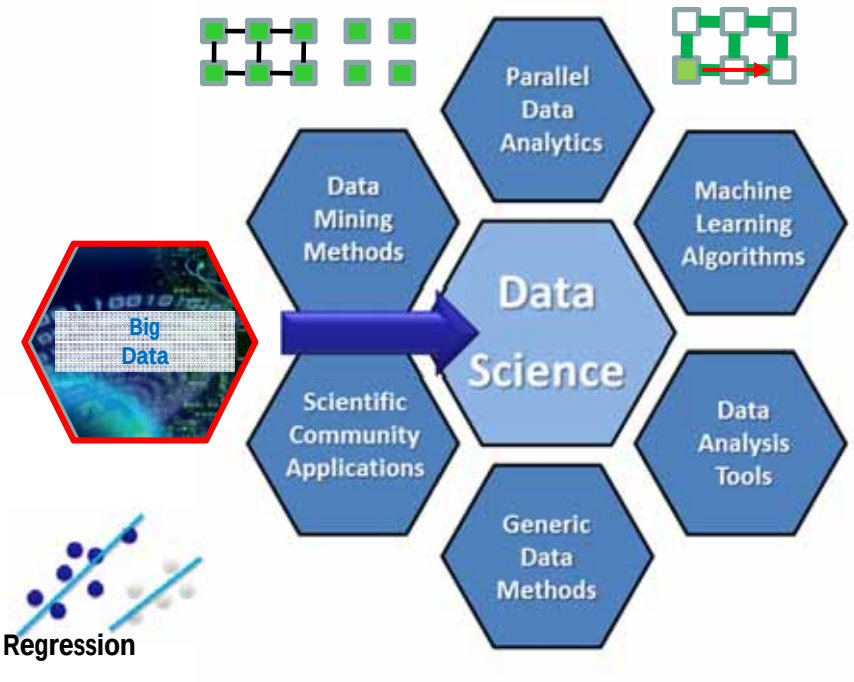


Juelich Supercomputing Centre Context

- **Commercial cloud infrastructures & offerings lack this needed structure for science**

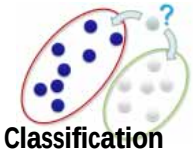


- Support data-intensive science and engineering applications
- Explore computing that is more intertwined with data analysis



Reliable Infrastructures Data Management, Security & Access

Clustering



Classification

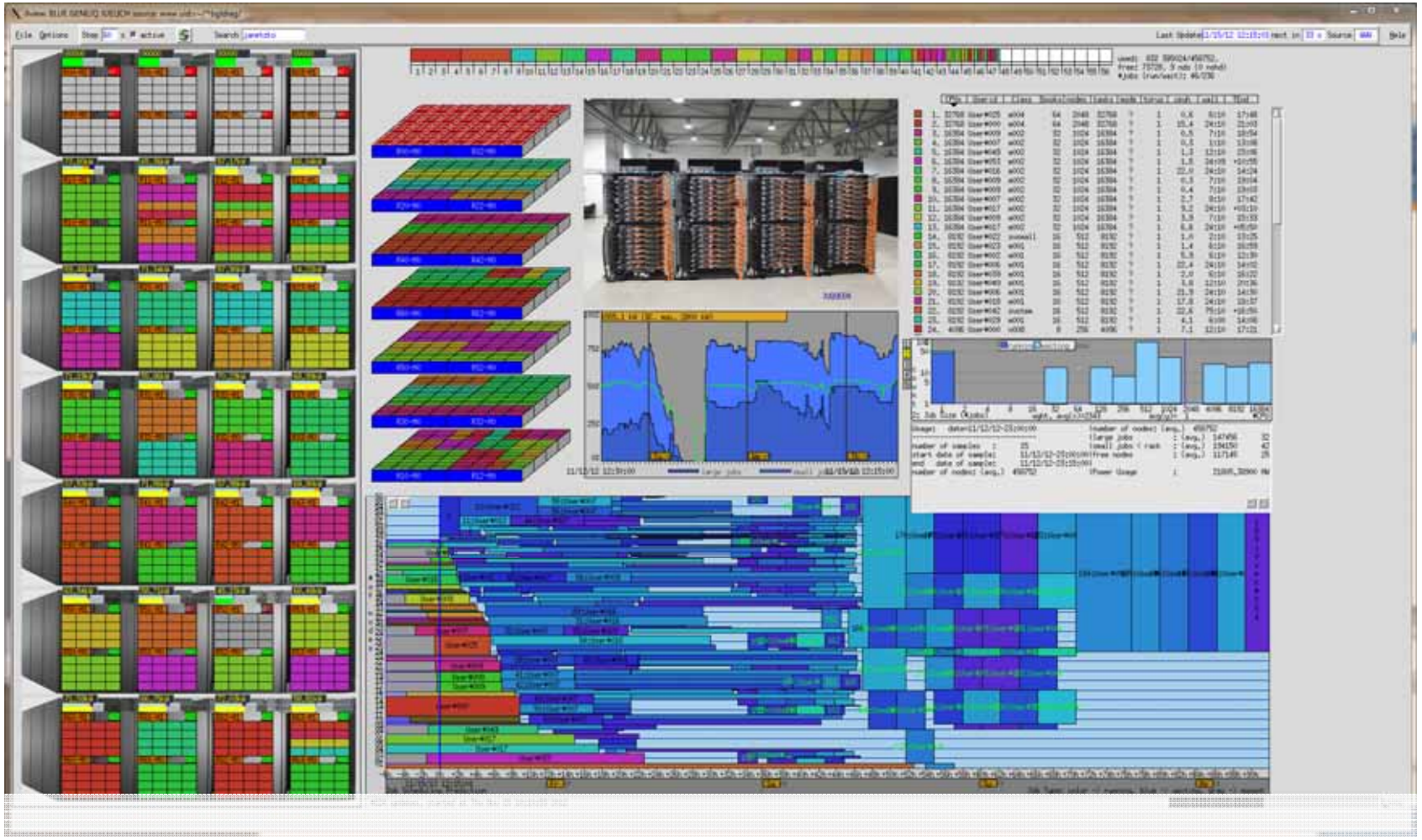
Regression

HPC Technology Advances: TOP 500 List (June 2014)

	Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)	
	1	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808	***power challenge***
CRAY	2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209	
IBM	3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890	
	4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660	
IBM	5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945	
CRAY	6	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Cray Inc.	115,984	6,271.0	7,788.9	2,325	***EU #1***
	7	Texas Advanced Computing Center/Univ. of Texas United States	**Stampede** - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P Dell	462,462	5,168.1	8,520.1	4,510	
IBM	8	Forschungszentrum Juelich (FZJ) Germany	**JUQUEEN** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	458,752	5,008.9	5,872.0	2,301	***EU #2***
IBM	9	DOE/NNSA/LLNL United States	**Vulcan** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	393,216	4,293.3	5,033.2	1,972	
CRAY	10	Government United States	Cray XC30, Intel Xeon E5-2697v2 12C 2.7GHz, Aries interconnect Cray Inc.	225,984	3,143.5	4,881.3		

[8] TOP 500 supercomputing sites

HPC Technology Advances – BlueGene/Q Example



1.6 GHz; 16 cores per node; 458,762 cores; 5,9 PF peak; 448 TByte main memory

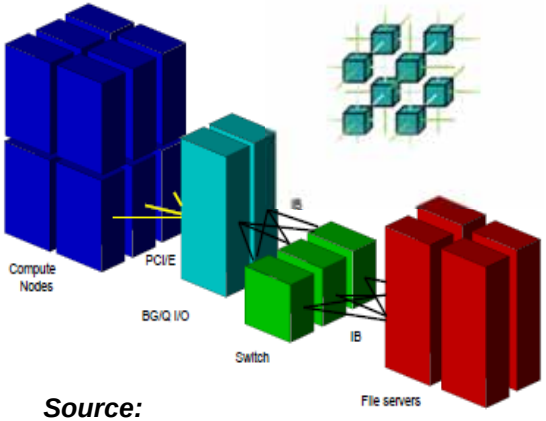
[9] LLView Tool

Selected HPC Technology Advances

Massively growing amount of cores/CPUs

- Reason: clock frequency of chips cannot be increased (hitting the ‘heat barrier‘ → ‘multi-core era‘)
- Impact: more precise & fine-granular simulation of reality becomes better (incl. real data validation)
- Effect: better simulation & prediction accuracies create increasingly ‘big data‘

Ever-increasing HPC technology capabilities



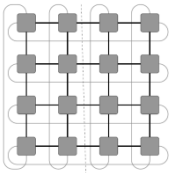
Source: IBM

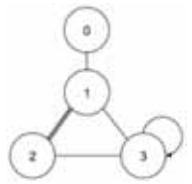
- Reason: application-oriented design
- Impact: more complex features of systems (e.g. network topologies mesh & 5D torus, different levels of caches, ‘multi-threading‘, I/O nodes, ...)
- Effect: Programming of heterogenous systems gets increasingly complex & vendor dependent

HPC Technology Advances – Network Topologies Example

Scalable, parallel I/O, fault tolerance, task-core mappings

E.g. What happens if one core out of millions fail?



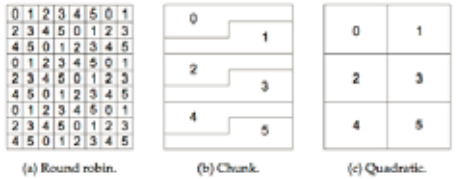


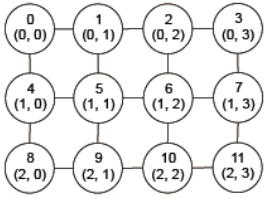
Execution units

i.e. processes

processing elements

i.e. CPUs



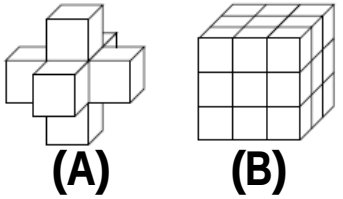


Work on Task-Core Mapping Algorithms

- NP hard problem – OS has no information about the application
- Users can enforce a specific core binding within the application
- Approach: Mapping model as graph; connection & distance matrix C_t and D_t
- Optimize the mapping: t (App) -> w (HW) based on an objective function
- Results: (A) 1-3%; (B) 1-2% application performance gain

$$E = \sum_{\{t_i,t_j\} \in E} C_t[t_i,t_j] D_w[m(t_i),m(t_j)]$$

Heatmap application example



➢ **Research activities with Stefan Klauck (master thesis) – HPI, University of Potsdam, Germany**

Web Technology Advances influence Science & Engineering

E.g. How can we hide the complexity of distributed & parallel systems?

Secure Web-based access

Hide complexity for users with Gateways

- Seamless Web-based & secure access
- Geographically distributed jobs & data
- Application domain-specific functionality

CIPRES Science Gateway

- Performs analysis that infer phylogenetic trees
- Uses user data-driven approaches

[10] CIPRES Science Gateway

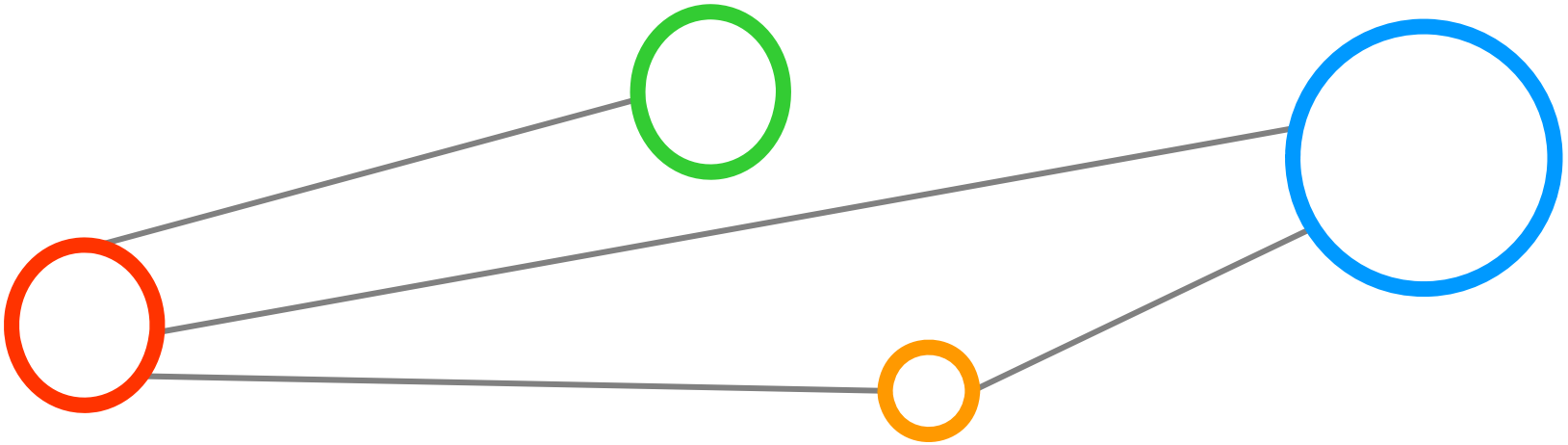
Ultrascan Science Gateway

- Bio-chemical data analysis
- Uses ultracentrifugal datasets

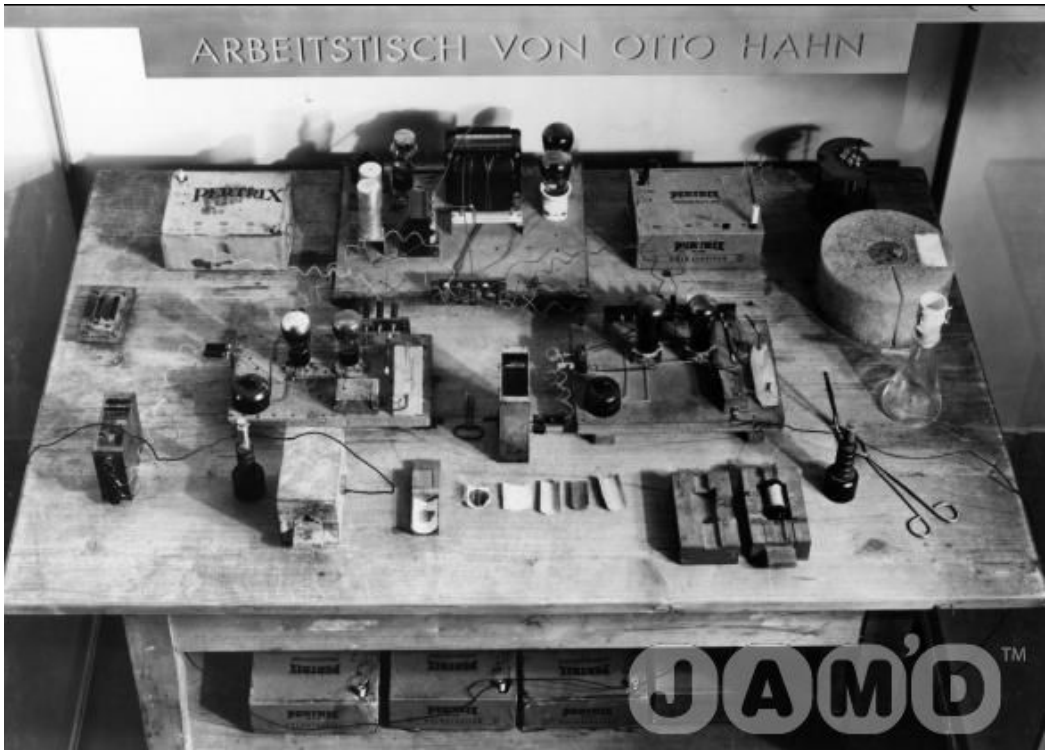
[11] S. Memon, M. Riedel, and B. Demeler et al., Concurrency & Computation: practice and Experience, DOI: 10.1002/cpe.3251, 2014

➢ **Research activities with Shahbaz & Shiraz Memon (PhD thesis) – Juelich Supercomputing Centre**

Parallel and Scalable Methods



‘Big Data‘ Applications – What are the right methods?



The right equipment and workbench… simple – yet powerful

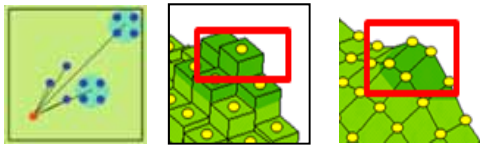
- Used by Otto Hahn (1879 – 1968) and Fritz Strassmann in December 1938
- No opportunity to use High Performance/Throughput Computing & Clouds
- Enabled to provide the first chemical evidence of nuclear fission products

Parallelization – Fundamentals

Two major reasons to engage in parallelization

- A single core is too slow to perform the required (automated) task(s)
  (e.g. in a certain constrained amount of time; effect on usability of systems)
- The available memory is not sufficient on a single system
  (e.g. to tackle a problem in a required granularity or precision)
- Reasons relevant for computational simulations and analysing ‘big datasets‘

Scalability to large numbers of cores



- Requires good domain decomposition techniques
- Hindered by the domination of serial parts in parallel applications
  (i.e. Amdahls Law; reasons: shared I/O, startup overhead, wrong algorithm,..)

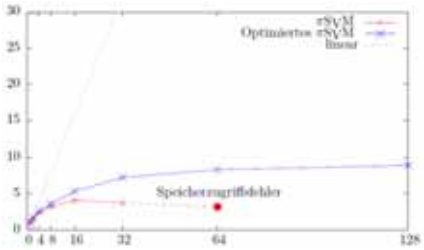
$$S_f = \frac{P_f^{\,P}}{P_f^{\,S}} = \frac{1}{S + \frac{1-s}{N}}$$

[12] G. Amdahl, Validity of the single processor approach to achieving large scale computing capabilities, 1967

Parallelization – Domain Decomposition Example

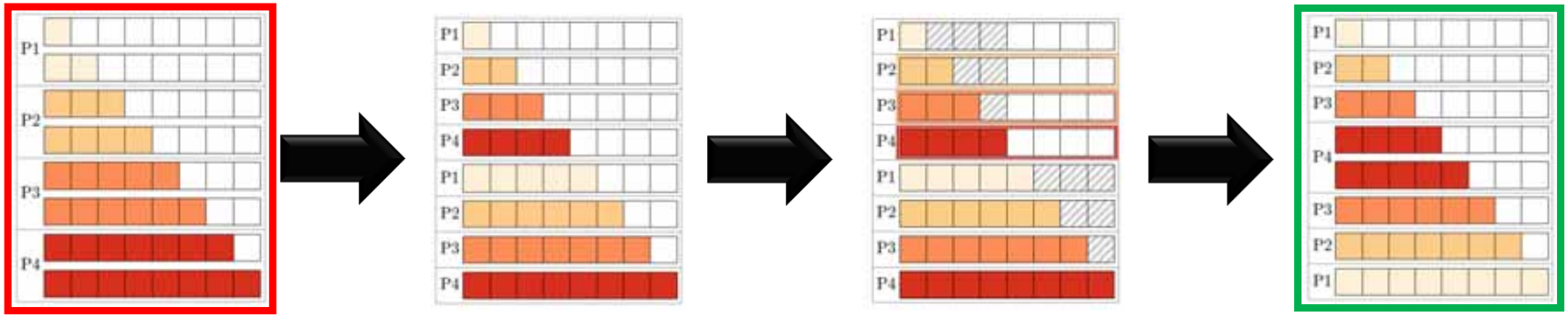
Scalable, parallel I/O, fault tolerance, task-core mappings

E.g. How do we distribute ‘big data‘ over cores & store to disks?



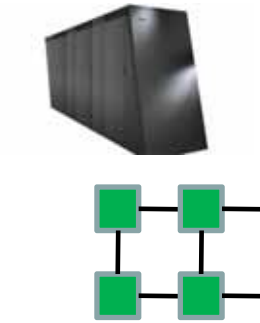
Work on Domain Decomposition Algorithms & ‘Load Imbalance‘ (e.g. Parallel Support Vector Machines application, etc.)

- Optimize existing algorithms, e.g. with parallel ‘collective operations‘
- Enable better parallelization with specific parallel datatypes (→ parallel I/O)
- Optimize scalability with different data distribution approaches



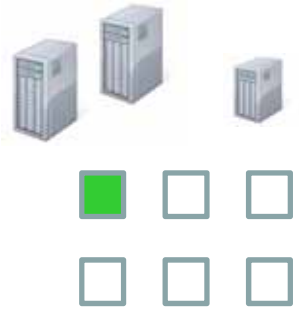
➢ **Research activities with Matthias Richerzhagen (bachelor thesis) – Juelich Supercomputing Centre**

Parallelization – Paradigms



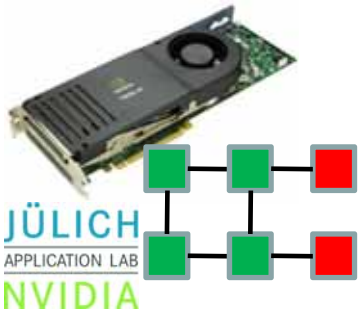
High Performance Computing (HPC)

- Massively parallel computing on large-scale supercomputers
- Good (costly) interconnects between nodes (e.g. Infinband)
- Dominant programming models: MPI, OpenMP, hybrid



High Throughput Computing (HTC)

- Nicely (embarassingly) parallel computing on small-scale
- Loosely coupled interconnects, internet (e.g. normal ethernet)
- Dominant programming models: task-oriented, farming, etc.



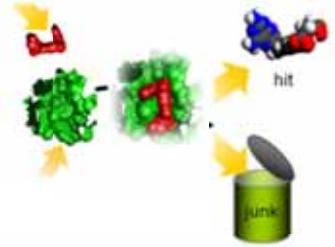
General Purpose Graphical Processing Units (GPGPUs)

- But very vendor-specific programming (CUDA, OpenCL)
- Emerging: Combination with previous two paradigms useful

Parallelization – Paradigms Interopability Applications

Computational simulation *& in-memory technologies*

E.g. How we can combine the strength of different paradigms?



Accelerate drug design simulations

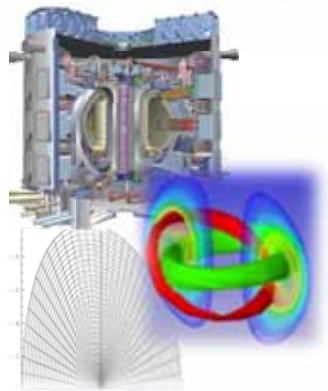
- HTC: molecular docking (Autodock)
- HPC: molecular dynamics (AMBER)

[13] M. Riedel & M. Hofmann-Apitius et.al , 2008

Virtual Physiological Human (VPH)

- HTC: HemeLB – low-scale pressure field
- HPC: HemeLB – high-scale velocity field

[14] M. Riedel, S. Zasada & P. Coveney et al., 2010



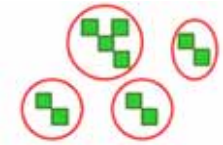
Fusion simulations for ITER Application

- HTC: HELENA high resolution code
- HPC: ILSA stability calculation

[15] S. Memon, M. Riedel, D. Tskhakaya and C. Konz et al. , 2010

Parallelization – Reliable Infrastructures to Satisfy Demands

Grid Computing Infrastructures (HTC-oriented)



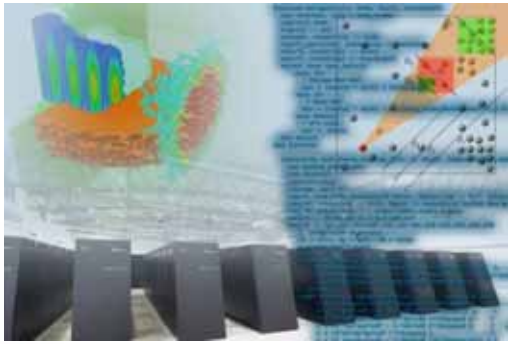
- Example: European Grid Infrastructure (EGI)

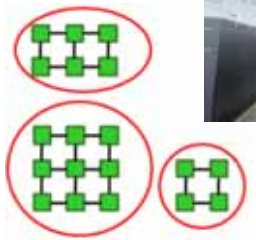
- **'Results today only possible due to extraordinary performance of Accelerators – Experiments – Grid computing'.**

[16] Rolf-Dieter Heuer, CERN Director General, in the context of the Higgs Boson Discovery

Supercomputing Infrastructures (HPC-oriented)

- Examples: Partnership for Advanced Computing in Europe (PRACE)





Hybrid Infrastructures (HPC and HTC-oriented, including visualization systems, etc.)

- Examples: Extreme Science and Engineering Discovery Environment (XSEDE)

Parallelization – US Hybrid Infrastructure XSEDE Example

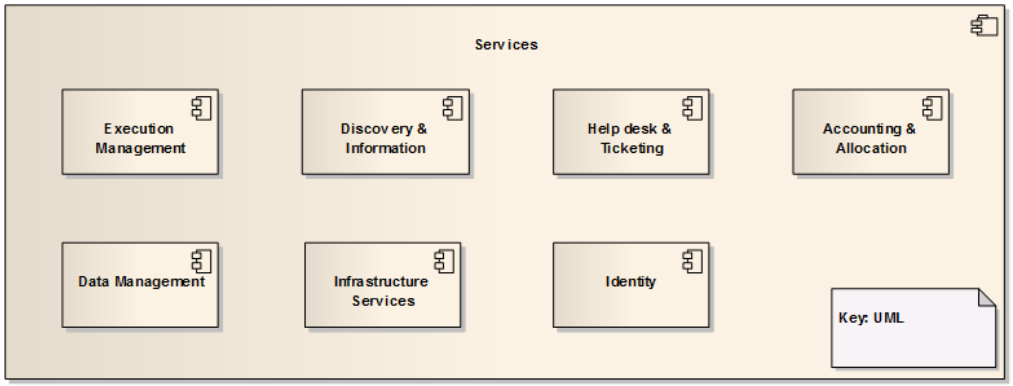
Secure, trusted & distributed infrastructures & archives

E.g. How we can separate architecture from implementation?

[17] F. Bachmann, I. Foster, A. Grimshaw, D. Lifka, M. Riedel, S. Tuecke, 'XSEDE Architecture Level 3 Decomposition', 2014

Next generation infrastructure

- Offers massive amounts of heterogenous resources
- Requires architectural design guided by real applications



Kraken @ NICS
(1.2 PF Cray XT5)
Stampede @ TACC
(7+PF Intel's MIC)
Gordon @ SDSC
(341 TF Appro SMP cluster)

Lonestar @ TACC
(302 TF Dell Cluster)
Trestles @ SDSC
(100 TF Appro Cluster)
Blacklight @ PSC
(36 TF SGI UV)

Future Grid

Map-Reduce
(Hadoop, Twister)

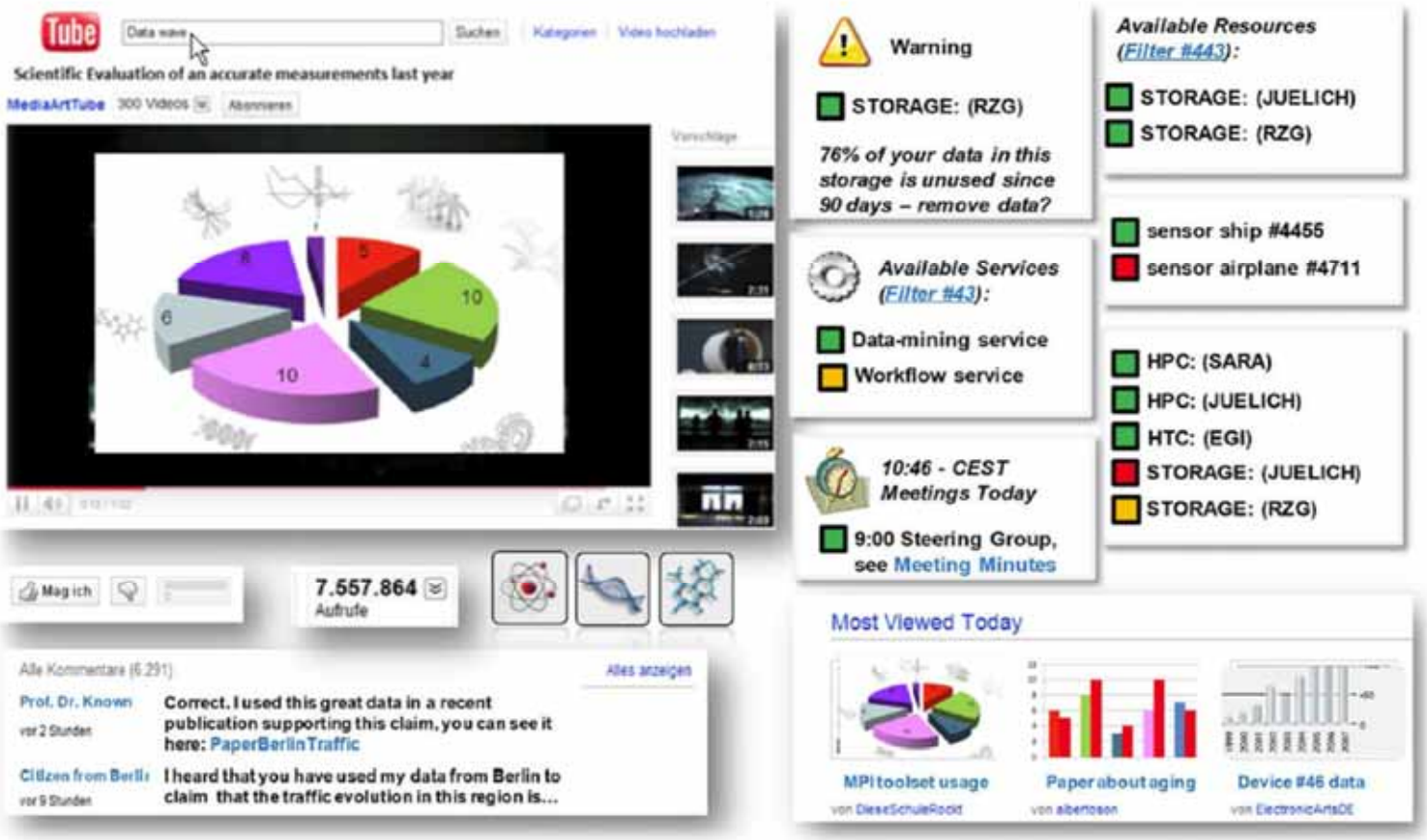
[18] UoIceland Teaching Project

Parallelization – Large-scale Data Infrastructures for Results

Secure, trusted & distributed infrastructures & archives

E.g. How we enable application reproducability and data sharing?

[19] M. Riedel and P. Wittenburg et al., Journal of Internet Applications, 2013



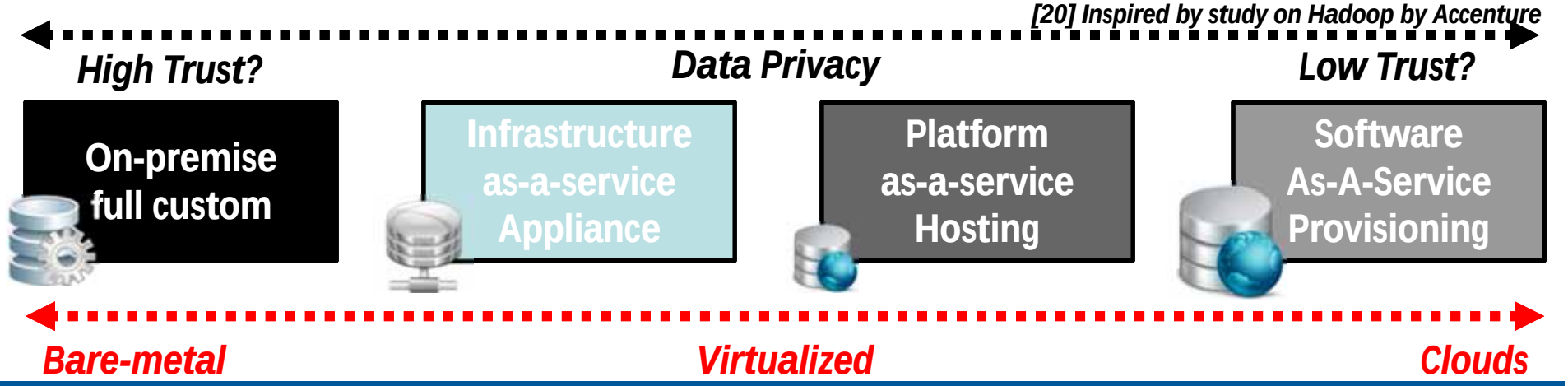
Parallelization – Cloud Computing Infrastructures

Pro

- Bring computation to the data (in case of ‘big data‘ safe transfer time/costs)
- Costs (not always) and maintenance overheads are significantly reduced
- Enable more simple parallelization methods (e.g. map-reduce vs. MPI)

Contra

- No domain-specific expertise alongside cloud providers as in large centers
- Scientific environments & grant process not prepared to ‘swipe credit card‘
- Tend to have more unstable technologies or switch quickly to new ones



Parallelization – Scalable Approaches & Tools for 'Big Data'

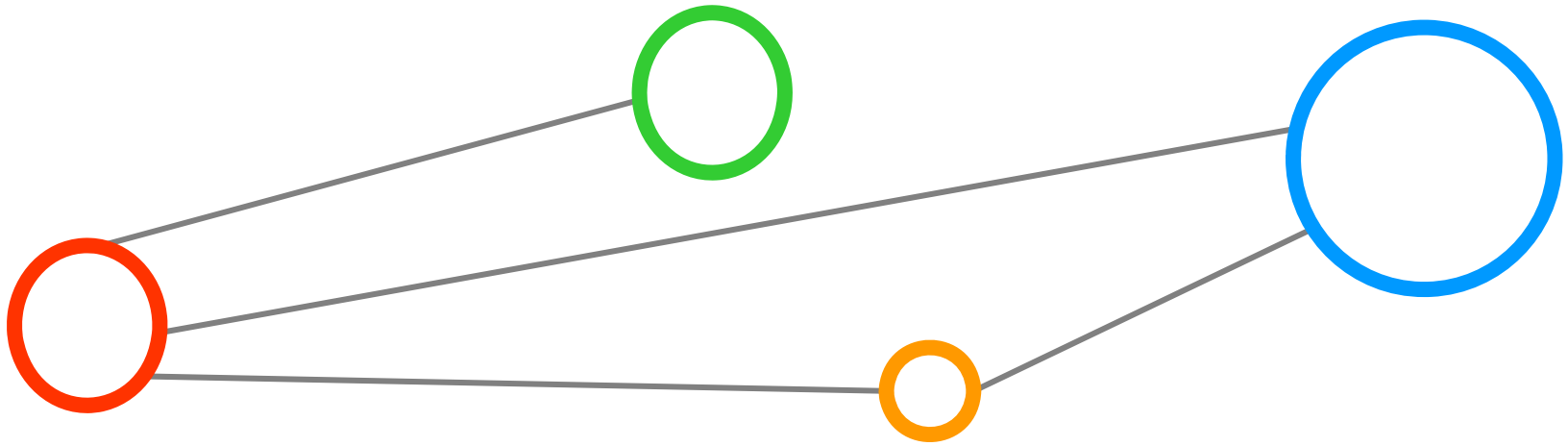
Many are available and promise much functionality

- Critical review although brings surprises (here survey focus on SVMs)

Tool	Platform Approach	Parallel Support Vector Machine
Apache Mahout	Java; Apache Hadoop 1.0 (map-reduce); HTC	No strategy for implementation (Website), serial SVM in code
Apache Spark/MLlib	Apache Spark; HTC	Only linear SVM; no multi-class implementation
Twister/ParallelSVM	Java; Apache Hadoop 1.0 (map-reduce); Twister (iterations), HTC	Much dependencies on other software: Hadoop, Messaging, etc.
Scikit-Learn	Python; HPC/HTC	Multi-class Implementations of SVM, but not fully parallelized
piSVM	C code; Message Passing Interface (MPI); HPC	Simple multi-class parallel SVM implementation outdated (~2011)
GPU accelerated LIBSVM	CUDA language	Multi-class parallel SVM, relatively hard to program, no std. (CUDA)
pSVM	C code; Message Passing Interface (MPI); HPC	Unstable beta, SVM implementation outdated (~2011)

- **The availability and functionality of open tools is limited and still need to increase**

Selected ‘Smart Data Analytics‘ Applications



Ultrascan Bio-Chemical Methods

Secure, trusted & distributed infrastructures & archives

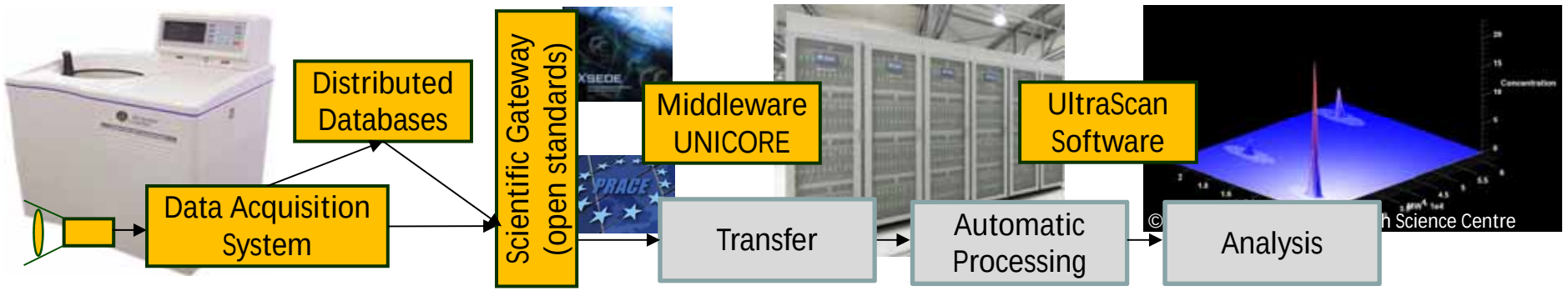
[11] S. Memon, M. Riedel, and B. Demeler et al., Concurrency & Computation: practice and Experience, DOI: 10.1002/cpe.3251, 2014

Statistical data mining

Task: Model data from hydrodynamic experiments to describe molecules under dynamic solution conditions [many users & computationally expensive inverse problems]

Analytical Ultracentrifugation Experiments

- Secure Web-based access for domain science
- Characterize molecules; molecular interaction
- Methods to eliminate noise, experiment & simulation data used by Ultrascan software



➢ **Research activities with Shahbaz & Shiraz Memon (PhD thesis) – Juelich Supercomputing Centre**

Location-based Social Network-based Health Analytics (1)

Statistical data mining

Interactive Visual Analytics

Towards ‘interactive visual analytics‘ using parallel methods

- Goal: Answer health related questions from publicly available data
- E.g. estimated emissions/region correlated with measurements stations
- E.g. estimated pollution/emissions breathing/person/region

Approach: Towards interactive data exploration

- Open data source: OpenStreet Maps (OSM) maps/streets
- „Click-free“ visualization in a browser, i.e. no GUI applications
- Support for typical overlays (density maps, polygon creation)

➢ **Research activities with Markus Goetz (PhD thesis) – Juelich Supercomputing Centre, UoIceland**

Location-based Social Network-based Health Analytics (2)

Scientific Domain Area

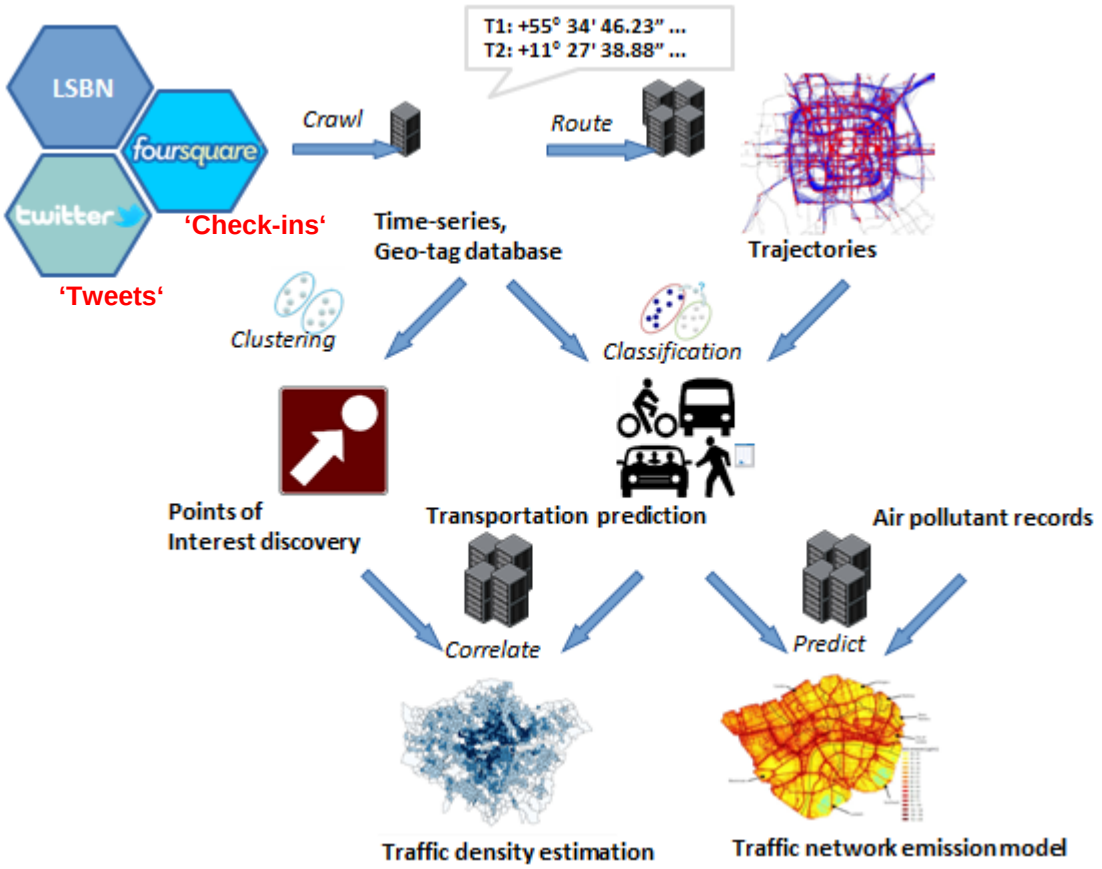
- Smart Cities approaches compined with Health Analytics Research

Scientific Outcome

- Traffic density estimation
- Network emission model

Location-based Social Networks (LBSN) Data

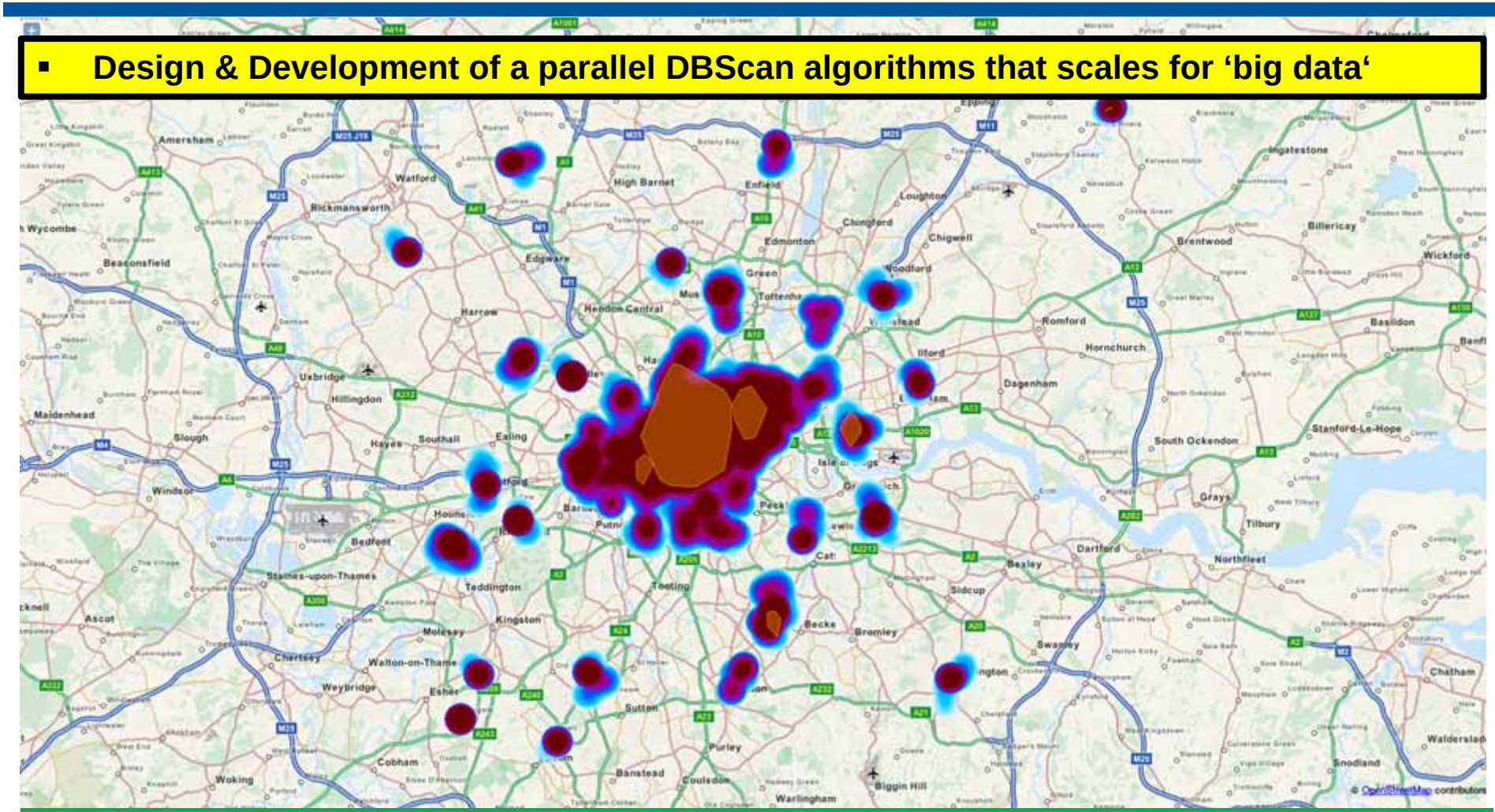
- Open data sources: Twitter & Foursquare
- Plan: Validation with real measurements in cities



➢ **Research activities with Markus Goetz (PhD thesis) – Juelich Supercomputing Centre, UoIceland**

Location-based Social Network-based Health Analytics (3)

- **Design & Development of a parallel DBScan algorithms that scales for ‘big data‘**



London Clusters – 6/1/2014, 1h time slice beginning at 18:00 UTC – using parallel Density-Based Spatial Clustering of Applications with Noise (DBScan)

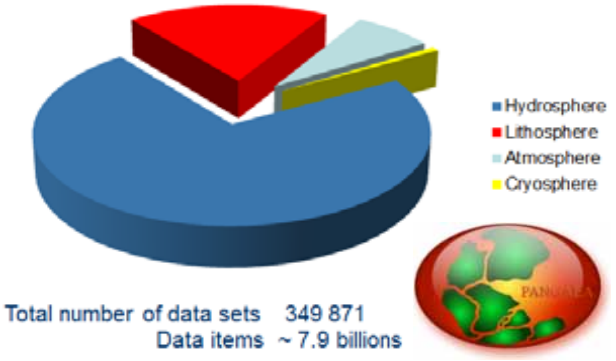
- **Research activities with Markus Goetz (PhD thesis) – Juelich Supercomputing Centre, UoIceland**

Automatic Outlier Detection of Large quantities of Data

Data of the PANGAEA data collection

- Massive volumes of data (‘underutilized‘)
- Find outliers – then check by domain scientists

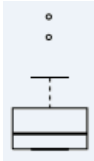
Statistical data mining



[21] PANGAEA data collection

Goal: Automatic outlier detection using parallel dbscan algorithm

- Better measurement devices produce orders of magnitudes more ‘big data‘
- Manual quality control becomes impossible and error-prone
- Automate the quality control process (→ parallelization)



e.g. boxplots not feasible

➢ **Research activities with Shiraz Memon (PhD thesis, Juelich) & Robert Huber (MARUM, Bremen)**

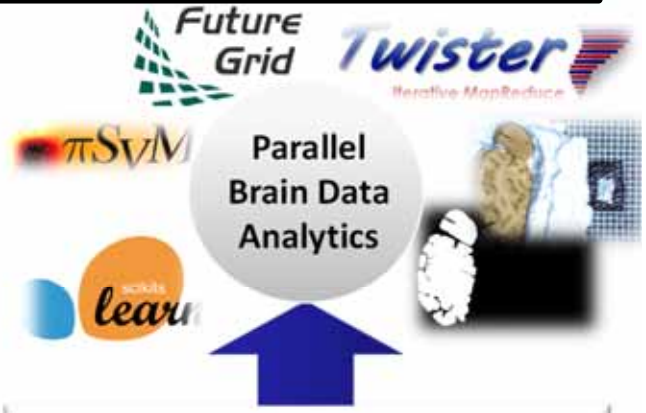
Classification Methods in Remote Sensing

- **Using methods like feature extraction, reduction, selection can reduce 'big data'**

Statistical data mining

Parallel smart analytics

[22] G. Cavallaro and M. Riedel et al. 'Smart Data Analytics Methods for Remote Sensing Applications', IGARSS 2014



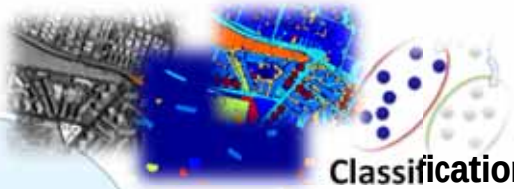
Sattelite Data (Quickbird)

Parallel Support Vector Machines (SVM)

πSVM

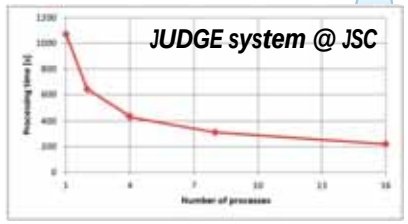
HPC & MPI

Classification Study of Land Cover Types



Classification++

'Best Practices'



„Reference Data Analytics" for reusability & learning

CRISP-DM Report	Openly Shared Datasets	Running Analytics Code

- **Research activities with Gabriele Cavallaro (PhD thesis, UoIceland) on Self Dual Attribute Profile**

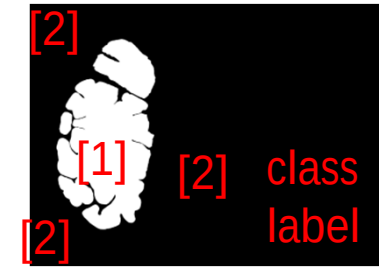
Techniques to Understand the Human Brain (1)

- **Build ‘reconstructed brain (one 3d volume) that matches with sections & block images**
- **Understanding the ‘sectioning of the brain’ and support automation of reconstruction**

Statistical data mining

Parallel smart analytics

1. Some ‘pattern‘ exists
   - Image content classification
2. No exact mathematical formula exists
   - No precise formula for ‘contour of the brain‘
3. Dataset (next: 5 brains, >100.000 pixels, 2PB raw)
   - Block face images (of frozen brain tissue)
   - Every 20 micron (cut size), resolution: 3272 x 2469
   - ~ 14 MB / RGB image
   - ~ 8 MB / corresponding mask image (‘groundtruth‘)
   - ~700 images → ~40 GB dataset



➢ **Research activities with Philipp Glock (Master thesis, Juelich) & Juelich Institute of Neuroscience**

Techniques to Understand the Human Brain (2)

Approach: Serial SVM implementations

- Data: sampled very small dataset, but equally balanced classes, 0.01%

Serial Scikit-learn (python) Example:

- Training time: ~39 minutes on JUDGE; Testing time: ~2 Min; Accuracy ~91%

Serial Matlab Example:

- Training time: ~3 hours on Laptop; Testing time: ~27 Min; Accuracy ~90,1%

Approach: Parallel SVM implementations

- Data: sampled very small dataset, but equally balanced classes
- No limit: theoretical scalable, but usefulness depends on datasets
- Incremented number of datasets, towards a sample size of 0.1%

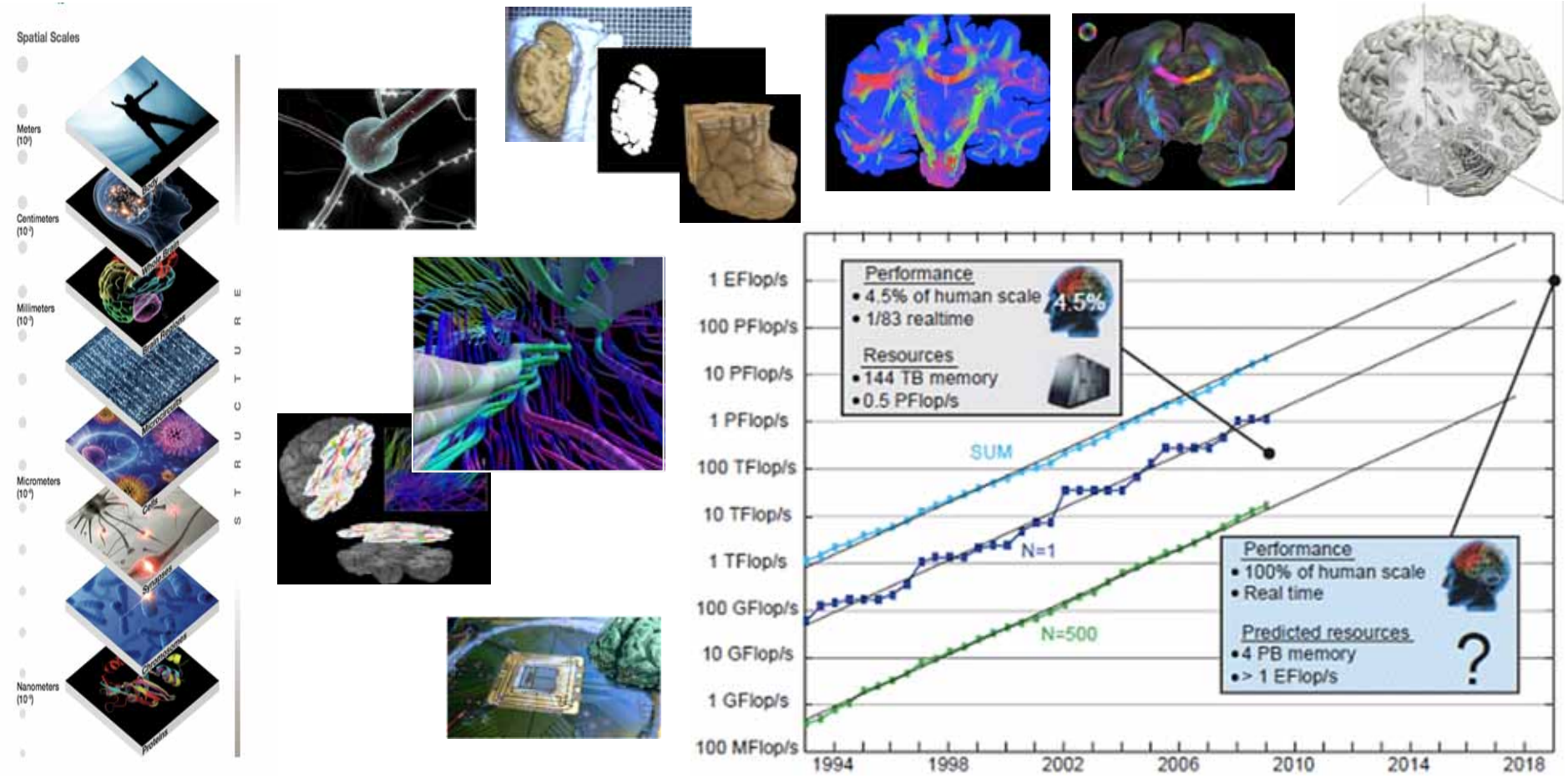
Twister (iterative Map-Reduce & Hadoop) Example:

- Training time: ~7 minutes on FutureGrid; Testing time: ~7 Min; Accuracy ~96%

➢ **Research activities with Philipp Glock (Master thesis, Juelich) & Juelich Institute of Neuroscience**

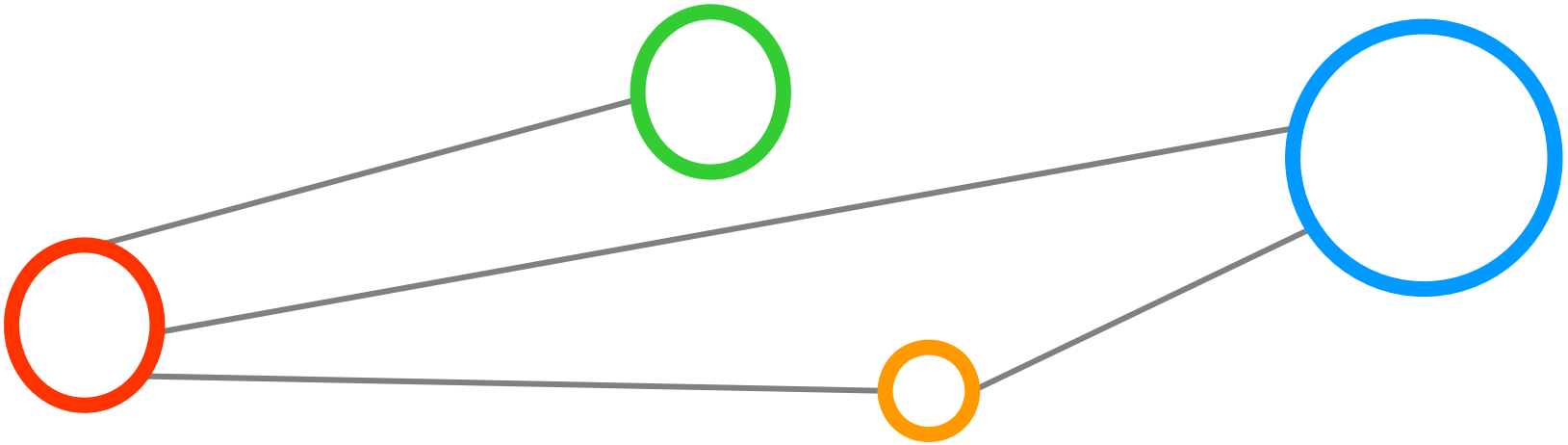
Techniques to Understand the Human Brain (3)

- **Computational and storage requirements raise demand for exascale technologies**



- **Research activities with Dr. Markus Axer & Dr. Timo Dickscheid (Post-Docs Juelich) & student teams**

Summary & References



Summary: Exascale Lighthouse with ‘In-Situ Analytics‘

Statistical data mining

Parallel smart analytics

NoSQL & key-value DBs

SQL Databases & storage

Scientific visualization

Computational steering

Interactive Visual Analytics

Secure Web-based access

Computational simulation & in-memory technologies

Scalable, parallel I/O, fault tolerance, task-core mappings

Secure, trusted & distributed infrastructures & archives

Smart Data Analytics:

Combination of these (often disjunct) Research Areas

Strategic Research towards a 2020 Vision

References (1)

[1] Riding the Wave, EC Report, 2010

[2] A Surfboard for riding the wave, Report, 2012

[3] DoE ASCAC Report, 2013

[4] Jeremy Ginsburg et al., 'Detecting influenza epidemics using search engine query data', Nature 457, 2009

[5] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani,
'The Parable of Google Flu: Traps in Big Data Analysis', Science Vol (343), 2014

[6] German Smart Data Innovation Lab Initiative, Website, Online: http://www.sdil.de/en/

[7] S. Joe Qin, 'Process Data Analytics in the Era of Big Data', AICHE Journal, 2014

[8] TOP 500 supercomputing sites, Online: http://www.top500.org/

[9] LLView Tool

[10] CIPRES Science Gateway

[11] S. Memon, M. Riedel, and B. Demeler et al.,'Advancements of the Ultrascan Scientific Gateway for Open Standards-based cyberinfrastructures', Concurrency & Computation: practice and Experience, DOI: 10.1002/cpe.3251, 2014

[12] G. Amdahl, Validity of the single processor approach to achieving large scale computing capabilities, 1967

[13] M. Riedel & M. Hofmann-Apitius et al., 'Improving e-Science with Interoperability of the e-Infrastructures EGEE and DEISA',
in Proceedings of the IEEE 31rd International Convention MIPRO 2008, Opatija, Croatia, pages 225–231, 2008

References (2)

[14] M. Riedel, S. Zasada & P. Coveney et al., 'Exploring the Potential of Using Multiple e-Science Infrastructures with Emerging Open Standards-based e-Health Research Tools', in Proceedings of the 10th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing (CCGrid 2010), Melbourne, Australia, pages 341– 348, 2010

[15] S. Memon, M. Riedel, D. Tskhakaya & C. Konz et al., 'Lessons Learned from Jointly Using HTC- and HPC-driven e-Science Infrastructures in Fusion Science', in Proceedings of 2nd IEEE International Conference on Information and Emerging Technologies (ICIET 2010), Karachi, Pakistan, 2010.

[16] Rolf-Dieter Heuer, CERN Director General, in the context of the Higgs Boson Discovery, YouTube Video, Online: http://www.youtube.com/watch?v=FgcoLUys3RY

[17] F. Bachmann, I. Foster, A. Grimshaw, D. Lifka, M. Riedel, S. Tuecke, 'XSEDE Architecture Level 3 Decomposition', 2014

[18] FutureGrid UoIceland Teaching Project, online: https://portal.futuregrid.org/projects/358

[19] M. Riedel and P. Wittenburg et al., 'A Data Infrastructure Reference Model with Applications: Towards Realization of a ScienceTube Vision with a Data Replication Service', Journal of Internet Applications, 2013

[20] Inspired by study on Hadoop by Accenture, Online: http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Hadoop-Deployment-Comparison-Study.pdf

[21] PANGAEA data collection, Online: http://www.pangaea.de/

[22] G. Cavallaro and M. Riedel et al. 'Smart Data Analytics Methods for Remote Sensing Applications', in proceedings of IGARSS 2014

Acknowledgements

Gabriele Cavallaro, University of Iceland

Tómas Philipp Runarsson, University of Iceland

Kristján Jonasson, University of Iceland

Stefan Klauck, Hasso-Plattner Institut, University of Potsdam

Markus Axer, Stefan Köhnen, Tim Hütz, Institute of Neuroscience & Medicine, Juelich

Selected Members of the Research Group on High Productivity Data Processing

Ahmed Shiraz Memon
Mohammad Shahbaz Memon
Markus Goetz
Christian Bodenstein
Philipp Glock
Matthias Richerzhagen

Thanks

Talk available at http://www.morrisriedel.de/talks